Hidden Resources – Strategies to Acquire and Exploit Potential Spoken Language Resources in National Archives

Jens Edlund, Joakim Gustafson
KTH Spech, Music and Hearing
Lindstedtsvägen 24, Stockholm, Sweden
E-mail: edlund@speech.kth.se, jocke@speech.kth.se

Abstract

In 2014, the Swedish government tasked a Swedish agency, The Swedish Post and Telecom Authority (PTS), with investigating how to best create and populate an infrastructure for spoken language resources (Ref N2014/2840/ITP). As a part of this work, the department of Speech, Music and Hearing at KTH Royal Institute of Technology have taken inventory of existing potential spoken language resources, mainly in Swedish national archives and other governmental or public institutions. In this position paper, key priorities, perspectives, and strategies that may be of general, rather than Swedish, interest are presented. We discuss broad types of potential spoken language resources available; to what extent these resources are free to use; and thirdly the main contribution: strategies to ensure the continuous acquisition of spoken language resources in a manner that facilitates speech and speech technology research.

Keywords: speech, national archives, oral history

1. Introduction

This position paper is based on the results of a government issued investigation into potential spoken language resources that are already in existence in Swedish national archives. We focus on key priorities, perspectives, and strategies that may be of general, rather than specifically Swedish, interest. After a brief background we begin with a survey of the broad types of potential spoken language resources that are available, then we move on and talk about the extent to which these resources are free to use and to what extent they are locked up in copyright or by other issues, and finally we end on the main contribution: strategies to ensure the continuous acquisition of spoken language resources in a manner that facilitates speech and speech technology research. The requirements in terms of structure, tools, and methods of maintenance for a spoken language resource infrastructure that can accept the resource influx and make it accessible for research and industry is however outside the scope of this paper.

2. Background

Over the past decade or so, several government issued investigations in Sweden have looked for a way to secure the availability of Swedish spoken language resources. The main motivations have been both to ensure a healthy and competitive development of Swedish speech technology and to ensure the availability of the speech technology components needed to provide accessible government, informational and educational materials. The resulting reports have, without exception, made clear the need for such resources, but have at the same time largely lacked viable suggestions as to how they should be acquired and made available (e.g. Ahrenborg et al., 2004; Andréasson et al., 2008; Borin et al., 2008; Språkrådet, 2012; A-focus, 2014), at least as far as speech resources are concerned – text data has often received more complete attention. In 2014, the Swedish government again tasked a Swedish agency, The Swedish Post and Telecom Authority (PTS), with investigating how to best create and populate an infrastructure for spoken language resources (Ref N2014/2840/ITP). As a part of this work, the department of Speech, Music and Hearing at KTH Royal Institute of Technology have investigated existing potential spoken language resources, mainly in Swedish national archives and other governmental or public institutions. This effort came to its end at the end of 2015, with a report forthcoming in 2016 (Edlund, 2016).

3. Related work

There are naturally a great number of projects, small and large, aiming to make use of existing language resources. The commercial giants take an interest, with Google/Alphabet being perhaps the most active in their use of the world's collected literature and of the huge amounts of spoken language present on for example their video sharing service Youtube. Large multinational companies such as Google, Apple, Amazon, Samsung and Microsoft also record enormous amounts of data from spoken user interactions, which give them an almost incomprehensible lead in data driven speech technology. Most of the data collected by these companies is not freely available to others.

Another set of actors are the large data distributors in the field, the Eurpoean Language Resources Association (ELRA) in Europe and the Linguistic Data Consortium (LDC) in the US. The services they provide often demands that someone has prepared the resources for research usage, and many of the resources they provide are not free.

There's also a number of projects aiming to gather either data or pointers and meta-data in order to make these more readily accessible. Examples that are relevant to Swedish include META-NET with META-NORD and CLARIN with SWE-CLARIN. These projects have often been successful as far as text resources are concerned, but fair less well when it comes to speech.

Other projects aiming to improve the development climate for speech research and speech technology (e.g. recently the CITIA/ROCKIT network) often point out the need for resources, but so far, a reasonably funded, long-term solution for speech resources is lacking.

4. Speech Data - The Hidden Resources

The progress of speech technology is changing the requirements on data for its further development. It is less dependent on data with a specific signal-to-noise ratio, specific recording conditions, and specific speakers behaving in specific manners than it was ten years ago; a much wider range of analyses is used; and a much wider range of data is interesting for speech researchers and other scholars to analyse. With a multitude of projects aiming to learn or understand speech in a manner similar to that of the infant – by "listening" to speech with little or no preconceptions – virtually any recording of speech constitutes useful data. Naturally, some data sets are more useful than others, or, at the very least, more obviously useful.

After a large number of interviews with representatives for Swedish governmental organizations – 3 agencies, 9 university departments, and 4 national archives directly, and another about two dozen agencies and cultural institutions with large collections through the coordinating government agency Digisam – the most remarkable insight is that virtually everyone has at least some speech data lying around. Discussions with some dozen representatives for foreign archives verify that this is the case in most, if not all, countries. In many cases, this is simply the result of some internal or external project from years back, often not digitized and rarely transcribed. An important lesson-learned is that finding out that it exists *works much better in person in face-to-face interviews than over email*, since there is often no obvious employee that is responsible for maintaining the data. E-mail, then, tend to go unanswered or be answered in the negative simple because they do not find their way to the right person. In other cases, such as for cultural heritage institutions or universities with a humanities programme, collection and maintenance of speech data can be a core function.

Apart from speech data that is collected for the sake of speech or speech technology research, there are a great number of functions that generate speech data as a side effect. Amongst the most notable: radio and television; governmental debate; (criminal) court sessions; contacts with care givers; interviews (for printed press, many governmental bodies produce internal magazines); social and ethnological studies; official conversations over the phone (these are sometimes recorded for quality assurance purposes) and user data when people speak to automated systems.

It should also be noted that stored text is not only generated by a far greater range of actions, but that it can be used for speech and speech technology research purposes, for example to build language models for automatic speech recognition.

5. Obstacles to Publishing - Will They Remain Hidden?

There are many reasons that most of the speech data that was not created explicitly for speech and speech technology research purposes is not available for such purposes.

5.1 Hidden data

Amongst the more obvious, some are relatively easy to get around: the existence of the data is often not known to people outside an inner circle that was part of its creation; the usefulness of the data is often not known to anyone at all. Others obstacles are considerably harder: the data might not be available in machine readable form; it may require equipment that is no longer available to read; or it may be misplaced or difficult to access physically.

5.2 Ownership

A much more complex set of obstacles has to do with legal issues. Ownership is often not easy to establish, and since it is a lawyer's duty to err on the side of caution, asking lawyers if it is possible to release to the public something with unclear ownership or copyright will often if not always meet with a negative response. The noticeable trend in Swedish organizations is to take a certain amount of risk: if the material is not sensitive, publish first and dealing with the problems when, and more importantly *if*, they occur. Jussi Karlgren, Adjunct Professor in language technology at KTH Royal Institute of Technology and founding partner of text analytics company Gavagai, has made this trend explicit in a call for data holders to "turn themselves into lightning rods" by releasing their data as a way to force into being a consistent and unambiguous copyright law where none exists.

5.3 Ethical considerations

Another layer of complication is added by ethical considerations, integrity, and the possibility of slander. Large quantities of speech data cannot reasonable be vetted for content, and publishing it without going through it first (which means listening to 100s or 1000s of hours of speech) comes with the risk that the contents violates somebody's rights. In many cases, the nature of the data (e.g. an interview about how to cook an omelette) makes this less likely, but in many cases the semantic contents of the data are largely unknown.

6. Strategies - Unearthing The Resources

A number of strategies for making these hidden resources at least in part available and useful have come out of the discussions with data holders, potential benefactors, interest organizations and speech technologists. Notably, the data holders are generally quite positive to the prospect of seeing their data come to use – this contrasts with other reports and previous investigations, and may be a sign of the times.

6.1 Derivate Data

Text resources such as the Swedish Språkbanken work around certain copyrights by transforming the data so that it cannot be used to access the original work (for which the copyright holds), for example by presenting text sentence by sentence in random order. Analogous methods can be used for speech technology. In the case of text, building N-gram models (with a reasonably low N) ensures that the original text cannot be rebuilt.

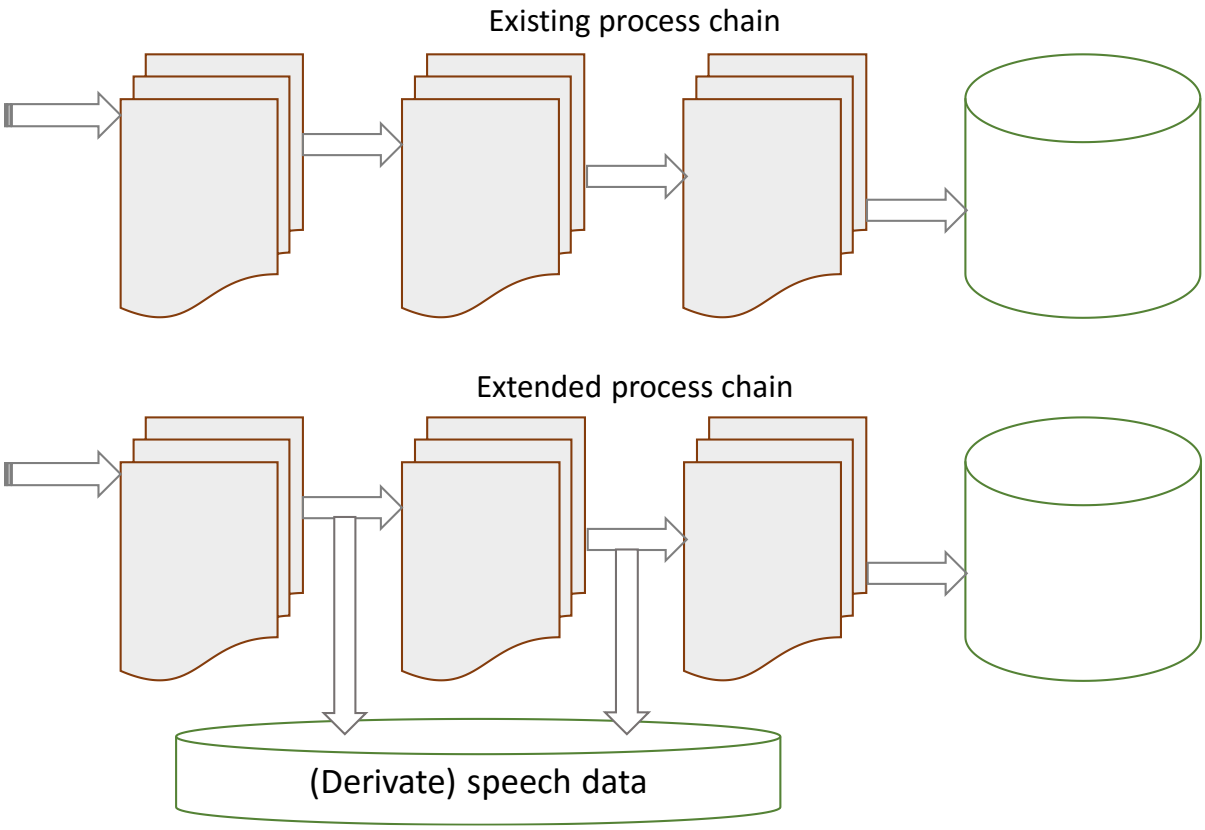

Figure 1: Illustration of process hooks. By inserting the hooks into processes rather than creating new processes that read data from disk and then pass it on to a speech data collection, the extra effort is "hidden" in the existing process chain.

Yet N-gram models, in particular if they can be assembled from a selection of contexts and domains based on metadata, are very useful in speech technology. As for the speech signal, in sensitive cases this can be parametrized in manners that hide the speakers' identities and still be useful, and various useful statistical models can be built without breaking copyright or revealing the semantic contents of the speech.

6.2 Process Hooks

Speech data has – compared to for example text – a large footprint. The effort it takes to access for example 100 hours of speech and transfer it from one organization to a spoken language resource is not negligible. This means that in most cases, specific funding and some kind of project format is needed for most acquisitions of speech data, even if it still exists in machine readable form. The effort required to set up such a project and get it accepted by all parties involved is daunting, unless we can turn it into a one-time effort that pays of over time. During discussions with several large Swedish resource holders, notably the National Library of Sweden (KB) and the Swedish Agency for Accessible Media (MTM), we have suggested to build in steps that tap the data into their existing production processes (Fig. 1).

The suggestions have met with enthusiasm, and we are now in the process of setting up process requirement for this strategy. The idea is that by building in data spoken language data collection into the production process, the main effort is a once-off undertaking, and from that on, any data produced in that process will more or less effortlessly populate the spoken language resource. One example of this is a language model tap built into text scanning processes, another is building acoustic models incrementally based on read texts.

6.3 Usage Data

Another method to dynamically and continuously populate a spoken language resource is to record the usage data from services provided. This is the method used by the large speech technology corporations, which has led to technological leaps in the case of standard question-answer speech recognition.

6.4 Donation

At least some publically available services, such as Google, have licenses that leave the users in possession of their own usage data. Setting up simple ways for users to donate such data would open one channel through which useful quantities of computer directed speech could be gathered. In a similar vein, services where people can help by transcribing or correcting speech resources could alleviate the cost of maintaining and refining speech data.

Several practical examples of crowd sourcing and human computation are being investigated and will be reported.

7. Recommendations

The methods above, once proofed – which is work we intend to do during 2016 – will not implement themselves. We suggest that by far the best way of efficiently creating an environment where contributing spoken language resources is a natural part of the everyday work of organizations that deal with speech in one way or another is by issuing recommendations, in much the same way that is often done concerning for example accessibility (a field that has much to gain from the existence of freely available speech resources). We propose to develop a set of best practices and guidelines for digitization, which includes steps that will ensure that language resources are tapped directly in the digitization process; best practices and guidelines for government agencies, which ensures that recordings of speech data take place in such a manner that the data is maximally useful as a spoken language resource without hampering the prime goal for recording the data; best practices for public tender and government procurement of speech technology and speech resources, ensuring that the resources bought as well as usage data is available for research and development; and best practices for EU and other research funding organizations to demand for example that data collected, at least derivatives, shall be made available at the end of a project.

8. Acknowledgements

This project was funded by The Swedish Post and Telecom Authority.

9. Bibliographical References

A-focus AB (2014). Nationell Språkresursbank – en utredning för Post- och Telestyrelsen. Report, Post- och Telestyrelsen.

Ahrenborg, L., Cooper, R., Josephson, O., Sågvall Hein, A. & Warnulf, B. (2004). Sverige behöver en strategi för språkteknologi. Letter to Ministers Marita Ulvskog, Leif Pagrotsky, and Thomas Östros, 28 januari 2004.

Andréasson, M., Borin, L. & Merkel, M. (2008). Habeas Corpus: A survey for SNK - a Swedish national corpus. Department of Swedish, University of Gothenburg.

Borin, L., Forsberg, M. & Lönngren, L. (2008). The hunting of the BLARK - SALDO, a freely available lexical database for Swedish language technology. Resourceful language technology. In *Festschrift in honor of Anna Sågvall Hein*, ed. by Joakim Nivre, Mats Dahllöf and Beáta Megyesi. Acta Universitatis Upsaliensis: Studia Linguistica Upsaliensia 7. 21 - 32.

Edlund, J. (2016). Nyttjande av offentligt tillgängliga svenska talresurser i en nationell talresursbank. Report, Post- och Telestyrelsen.

Språkrådet (2012). Infrastruktur för språken i Sverige – Förslag till nationell språkinfrastruktur för det digitala samhället. Report.